Zhang et al. EURASIP Journal on Wireless Communications and Networking (2016) 2016:217
DOI 10.1186/s13638-016-0709-y

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Model updating mechanism of concept drift detection in data stream based on classifier pool

Baoju Zhang, Lei Xue, Wei Wang*, Shan Qin and Dan Wang

Abstract

In data stream, concept drift often occurs in an unpredictable way, the classifier model which was learned from previous data is not accurate to the current data, so regular updating of the model is necessary. Moreover, updating too frequently will cause a negative impact on the clustering accuracy and the analysis of subsequent data. This paper proposed a model-updating mechanism for data streams with concept drift, and the updating mechanism is based on the calculation of correlation coefficient and information entropy. In terms of cumulative sum of entropy values and the predefined threshold, the system determined whether to update the model. After the model updating, classifier pool saves the existing model, and each classifier is used to describe a previous concept so that the system can detect recurring concepts drift and reduce the updating frequency.

Keywords: Data stream, Concept drift, Correlation coefficient, Information entropy, Classifier pool

1 Introduction

In many real problems, concepts often change over time and the system may not be able to possibly store all the data. Often, these changes make the model, which is built on old data, unable to adapt to the new data, and regular updating of the model is necessary. Clustering problems and learning the model from data become more challenging in the presence of concept drift. An effective learner should be able to track such changes and to quickly adapt to them. The exploration of concept drift detection method and the establishment of the system updating mechanism are topics of growing interest in the field of data mining research.

Handling concept drift has received a lot of attention in recent years. Actually, many updating mechanism for concept drifting data streams have been proposed. The authors in [1] proposed CVFDT algorithm, which could update its subtree and adapt to the new data environment by building the alternative subtrees. In [2], IncreDB2 algorithm detects the local concept drift real time, makes corresponding adjustment adaptively, and solves the problem of local concept drift well by updating the local drifted nodes. In [3], by dynamically changing its model to adapt to the target concept, M_ID4 algorithm has a good performance of accuracy and adaptability in a small number of training samples abrupt concept drift. Aboalsamh [4] proposed an incremental learning way and overcame the effect of concept drift phenomenon through accelerating the rate of update and updating the model continuously. The authors in [5] proposed a variable sliding window model; when the concept drift occurs, the sliding window can adjust the size of the window automatically. But the relevant work above exist a problem that the system updates the model over-frequently, and it would cause the reduction of clustering effect and excessive occupancy of resources.

There also are lots of studies on reducing the updating frequency in handling concept drift. Morshedlou and Barforoush [6] proposed an approach that uses the mean value and standard deviation to calculate and obtain the next concept probability; if a drift is detected and its probability is more than a threshold, the algorithm decides to behave proactively. The setting of a threshold is time consuming, and this algorithm only supports the numeric data sets. Katakis et al. [7]

* Correspondence: weiwangvip@163.com
Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

proposed that distance of recent vector in the pool is less than a predefined threshold, the classifier will update correspondingly by instances of the lately window, or a new model is saved. The setting of a threshold parameter is its weak point.

In this paper, first, we use a mathematical model that combined correlation coefficient with entropy to construct the updating model and preset threshold to avoid excessively frequent updates ultimately. Second, we extract the features of data and save the classifier model to the pool, and when new data is coming, the system chooses the most appropriate model or constructs a new classifier that suits the received data. The rest of this paper is organized as follows. In Sections 2 and 3, the proposed updating mechanism based on the classifier pool is present. Section 4 contains the experimental evaluations of the proposed algorithm. Section 5 concludes the paper.

2 Proposed updating mechanism

2.1 Correlation coefficient

The correlation coefficient, which is also referred to as Pearson correlation coefficient, describes the relationship between the two equal interval variables [8]. From the view of geometry, the correlation coefficient is the included angle cosine. From this point of view, the space corresponding to the random variables is composed of the equivalence class which is invariant under translation, and in such space, the standard deviation is the norm of vector, and the covariance is the dot product of the vector. According to the experience we learn from the plane, the higher the absolute value of two included angle cosine is, the closer the collinear two vectors are, and it also explains the meaning of related coefficient: in two groups of data, the higher the absolute value of correlation coefficient ρ_{XY} is, the more linear the two data sets are. The correlation coefficient can be expressed by:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \quad (1)$$

Correlation coefficient ρ_{XY}'s values are between −1 to 1, so when $\rho_{XY} = 0$, it is called irrelevant and when $|\rho_{XY}| = 1$, it is called complete correlation. There has linear functional relationship between X and Y; when $|\rho_{XY}| < 1$, the change of X value will cause the change of Y value, and the higher the absolute value of ρ_{XY} is, the bigger the change of Y will be. When ρ_{XY}'s absolute values are between 0.8 and 1, we call it strong correlation. We take the logarithm of correlation coefficient and obtain its trend diagram; as shown in Fig. 1, the x-axis denotes the correlation coefficient and y-axis denotes its logarithmic value.

2.2 Information entropy

We assumed that in a probabilistic system, there are n events, X_1, X_2,...,X_i,...,X_n, the probability of X_i event is $P_i(i = 1, 2, 3)$, and when the event X_i is generated, the given amount of information is $H_i = -\log_2 P_i$, in bit. For a probabilistic system consisting of N events, the average amount of information generated is:

$$E = -\frac{1}{n}\sum_{i=1}^{n} p_i \ln(p_i) \quad (2)$$

We called it information entropy, hereafter referred to as entropy [9].

Shannon introduced the concept of entropy to the information theory, to describe the uncertainty of a system structure. The bigger the entropy value is, the less the uncertainty is, whereas the lower the information entropy there is in the system, the more order the system is. Therefore, we can use the entropy to weigh the

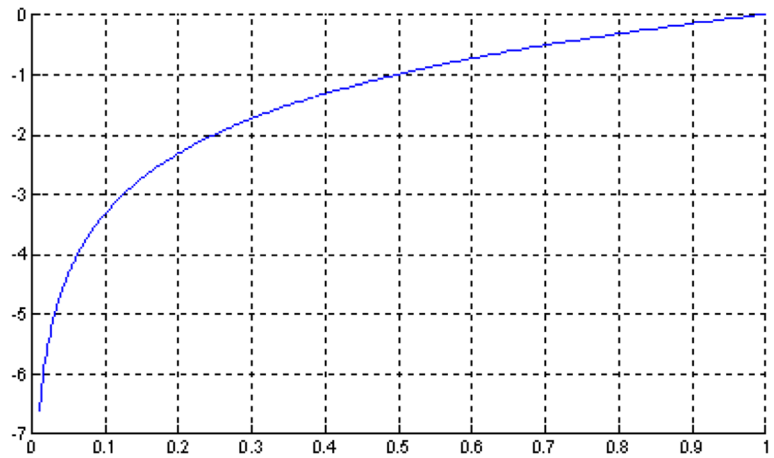

Fig. 1 The trend of the value of correlation coefficient after taking logarithm

degree of ordering in a system. In order to observe the distribution shape of entropy value H, which will change over time due to the change of the probability P that event X_1 occurs, we can depict the variation curve on Matlab platform. The x-axis and the y-axis denote the value of correlation coefficient and its corresponding entropy respectively, as shown in Fig. 2.

2.3 Updating mechanism based on correlation coefficient and entropy

In this paper, we use correlation coefficient to express the probability in the expression of entropy and give practical significance to the mathematical formula. We calculate the correlation coefficient of new data block and the original data block, and determine whether to update according to the calculative summation of entropy, we use *p to denote the calculative sum*. If the correlation coefficient is close to 1, and entropy will have a smaller value, there exists strong correlation; if the correlation coefficient is close to 0, and entropy will have a high value, there exists weaker correlation.

2.3.1 Data processing

The main function of the data stream initialization part is to process data of dynamic data stream and convert it to the form of static data block, so then the data can be processed by the system.

We supposed that a data stream can be partitioned into N data blocks and used D_1, D_2, D_3,..., D_N to express those blocks based on their order of arrival. Concept drifting means the concept in data has changed, and the overall distribution also has changed, assuming that probability distribution between arbitrary blocks is independent, and concept drift can be captured by measuring the degree of distributions' similarity between data blocks.

2.3.2 The establishment of mathematical model

In this paper, we combined entropy with correlation coefficient and regarded it as an index to determine whether the drifting occurs or not, finally accomplishing the aim of detecting concept drift. The correlation coefficient could measure the similarity and can be obtained from formula (3).

$$\rho_{D(1,i)} = \frac{\sum_{j=1}^{n}\left(x_j-\overline{x}\right)\left(y_j-\overline{y}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_j-\overline{x}\right)^2}\sqrt{\sum_{j=1}^{n}\left(y_j-\overline{y}\right)^2}} \quad (3)$$

Regarding the correlation coefficient as the probability and then substituting it to the expression in here, we can obtain the entropy value. The incoming data stream is partitioned into sequential blocks, and by calculating the correlation coefficient between the old and new data block, we can obtain the entropy value, and then accumulating the entropy, finally, get the mean entropy value $\bar{E}$.

$$\overline{E} = -\frac{1}{N}\sum_{i=1}^{N}\rho_{(1,i)}\ln\left(\rho_{(1,i)}\right) \quad (4)$$

2.3.3 Updating model

We introduced the concept of the classifier pool and saved models and concepts appeared previously in it, from which we can know if the concept newly presented exists or not in the data stream before. In this case, new data block exists in the following situations.

In the condition that concept drift have not occur in the data stream, there exists strong correlation between the previous and new block, the correlation coefficient is close to 1, the entropy value is small, and the module continues to use the original pattern.

On the condition that concept drift has occurred, we introduced the threshold in order to avoid the

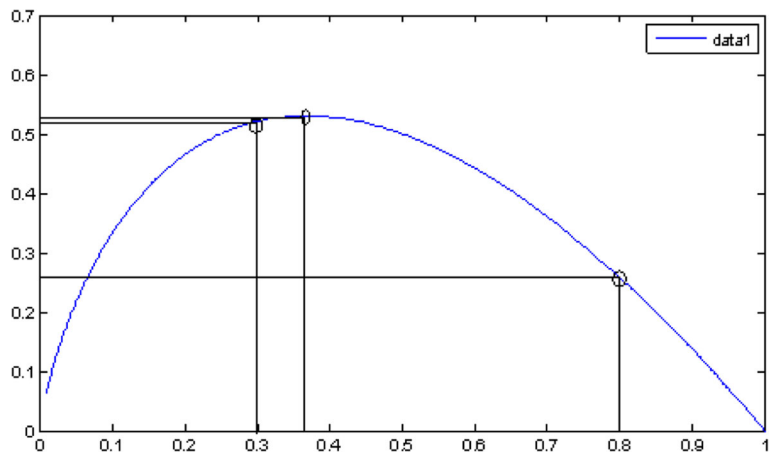

Fig. 2 The trend of entropy value within the scope

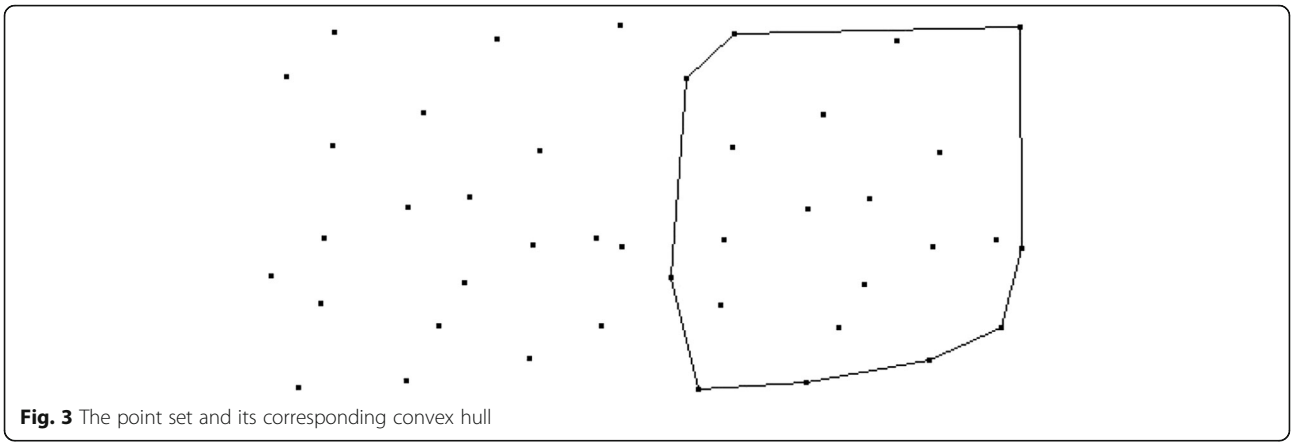

Fig. 3 The point set and its corresponding convex hull

model frequent updating. After the concept drift is being detected, we get the similarity comparison based on mean entropy value with history data in the pool and make the judgment on updating the model by comparing the mean entropy value $\bar{E}$ with presetting threshold ∂.

The mean entropy value $\bar{E}$ is calculated by Eq. (4); as a measure to detect the concept drift, we compare it with the threshold. As for the setting of the threshold, without an absolute boundary, it is hard to find a precise indicator to quantificate the threshold, and we choose an appropriate value from an interval as the threshold according to the experimental results.

1) If $\bar{E} < \partial$, there is a certain correlation between the old and new data blocks and the correlation is weak, and this proves that there have been similar concepts in the historical data, and not update. Although the change of concept is slight, through the accumulation of entropy, the model continues to detect the concept drift phenomenon for data stream clustering.
2) If $\bar{E} > \partial$, this proves that there have no similar concepts in the historical data, and the concept in data block is completely different from those in the pool. And in this case, we use the existing data block to replace the original data block and deposit the shape and characteristics of midpoint of the existing data block into the classifier pool, so as to re-training the model by using the new block.

3 Classifier pool

3.1 Shape extraction of data block: convex hull

Concept drift may be expected to recur in many practical applications, and concept features could be saved to the classifier pool so that they could be detected and reused later and achieve adapting more quickly to concept drift.

Convex hull is a minimum convex set which contains all the points in a point set [10]. By extracting shape features, we can get the shape diagram of the data block and save its center point. In this paper, we use convex hull to describe the shape feature of data block. In terms of convex hull's change between history data and current

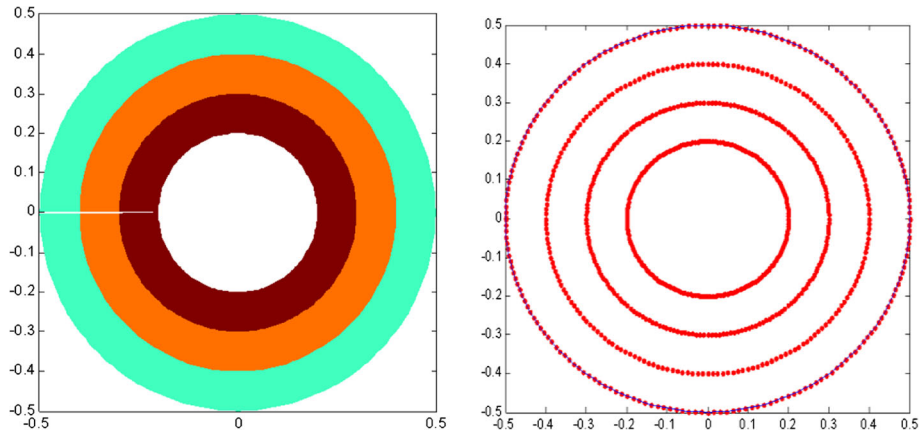

Fig. 4 The graph and its shape graph of circular data

data, potential drift could be distinguished. The convex hull is found by Graham algorithm, which mainly includes the following steps:

Algorithm 1 Graham algorithm

1: Select the point P_0 with the smallest ordinate value in S

2: Points in S sorted by polar angle , and get (P_1,P_2,...,P_n)

3: Push P_1

4: for i=2:n do

5: while the sorted polar angle form a simple polygon, and delete its concave point

6: pop P_{T2}, the second top point

7: Push P_i, the top point

In an arbitrary planar point set, after sorting and scanning, its convex hull is shown as Figs. 3 and 4 show the circular data and its convex hull.

3.2 The working mechanism of classifier pool

We introduced the classifier pool in this paper. By saving the classifier which corresponds to different concepts, the system model can choose the appropriate classifier from the classifier pool directly and has no need to retrain when recurring concept appears again.

A classifier has two attributes: shape feature and center point. By extracting the shape feature of data block and saving its center point, when the concept drift was captured, the system trained the data block and the model would not update if the new edge feature and center are similar with the original model. We would update the model and save it to the classifier pool if the edge shape and center changed obviously with the previous model. The system would use the previous model to process. In the data blocks which followed, the system invokes the previous model if it has a similarity with the saved patterns. Through the management of the classifier pool, the system achieves the detection of concept drift more efficiently.

1) If concept drift that occurred has been detected, the system will not update its model until the model cannot cover accumulative extent of the concept drift; instead, it will integrate the new coming data blocks, train those data, and get the new clustering model. Then, we compare the new model with the existing models in the classifier pool and get their similarities. The calculate method of similarity is based on KL-distance algorithm [11], and through calculation, we choose the nearest approximate as the new model.

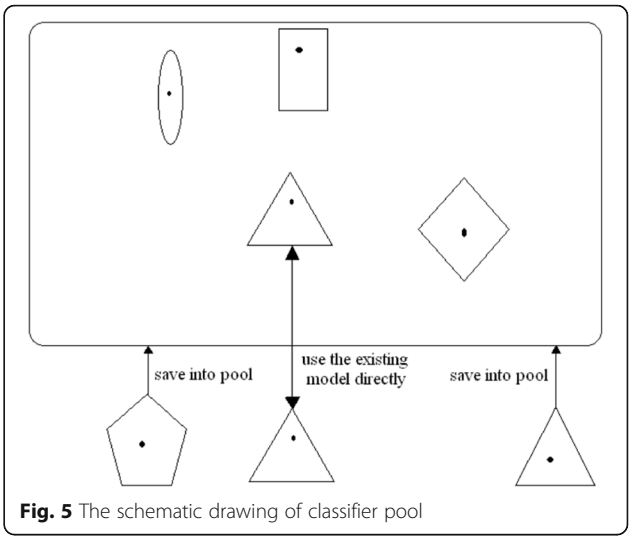

Fig. 5 The schematic drawing of classifier pool

2) If the models in the pool cannot match the new model, that is to say, the newly data block belongs to a new concept, there is no corresponding model in the pool that could cluster the new block. Therefore, we use it to train the new model and save it into the pool.
3) If there is no concept drift that occurred, then we use the nearest classifier to process new data blocks.

Its schematic drawing is shown in Fig. 5.

4 Experimental results and analysis

In this paper, we use power supply data set to verify validity and feasibility of algorithm. The data set is sampled from the grid data set within the main grid data and sub-grid data. Features of the data include hourly power supply of an Italian power company. The concept drifting in the data flow is mainly driven by the change over

Table 1 The correlation coefficient and entropy value of two data blocks

Time	Correlation coefficient	Entropy value	System model
0, 1	0.9284	0.0995	No concept drift
0, 2	0.9089	0.1253	No concept drift
0, 5	0.8755	0.1680	No concept drift
0, 6	0.7081	0.3526	Concept drift, accumulating
0, 7	0.7953	0.2627	Concept drift, accumulating
0, 8	0.5283	0.4864	Concept drift, accumulating
0, 10	0.6445	0.4085	Concept drift, accumulating
10, 13	0.9939	0.0087	No concept drift
3, 21	0.1573	0.4197	Concept drift

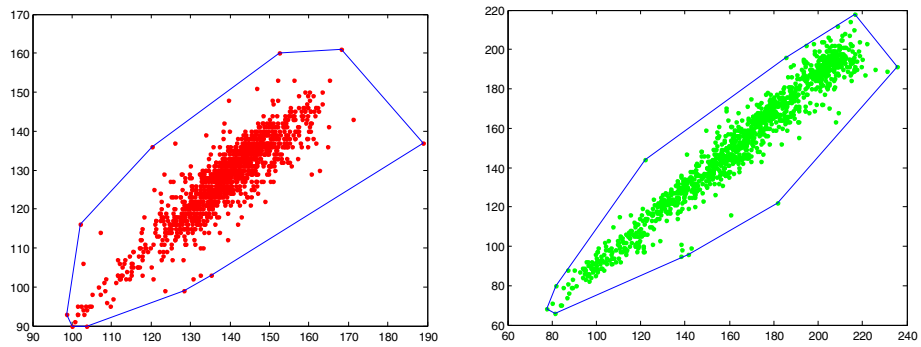

Fig. 6 The shape graph of 0 am and 8 am. **a** The grid data at 0 am. **b** The grid data at 8 am

time. It contains 1247 samples per hour, and the data is estimated from 0 am to 23 pm.

The concept drift caused by hours of a day is the main concern in our experiment, and the power supply needs to change accordingly depending on the time of day. As a result, we use the different data blocks of different time of a day to analyze the occurrence of the concept drift and decide when to update the model.

As shown in Table 1, we set 0 am's data set as the original data block, and the correlation coefficient of 0 am~5 am are close to 1, during which concept drift did not occur. We can know that concept drift were detected from 6 o'clock to 10 o'clock. Comparing data block of 13 am with those of 10 am, they have roughly similar energy consumption, correlation coefficient is close to 1, and there are no concept drifts that occur. Comparing data block of 21 pm with those of 3 am, the correlation coefficient is close to 0, the corresponding entropy has a larger value, and there exists concept drift.

In the experimental figure, we apply a two-dimensional figure, which consists of main grid data and sub-grid data, to reveal the distribution of power supply data, where x-axis represents main grid data and y-axis represents the sub-grid data. The shape graph of the power supply data at 0 am is shown in Fig. 6a, the concept drifts inside the block are detected at 6 am, the system choose to accumulate entropy values and not to update model immediately, before the drifting occurred at 6 am, and the mean entropy value is 0.1309. Concept drift was detected that occurs between 6 and 10 o'clock, the mean entropy value is 0.37755, and the threshold we preset should be between the two entropy values. In this paper, we set the value to 0.19. Therefore, the model accumulates the data of 6~8 am, and as shown in Fig. 7, we compare it with the data block of 0 am, and updates itself at 8 am. The comparison of pre- and post-updating models is shown in Fig. 8.

According to the shape feature in the classifier pool, it can be concluded that the new coming data block

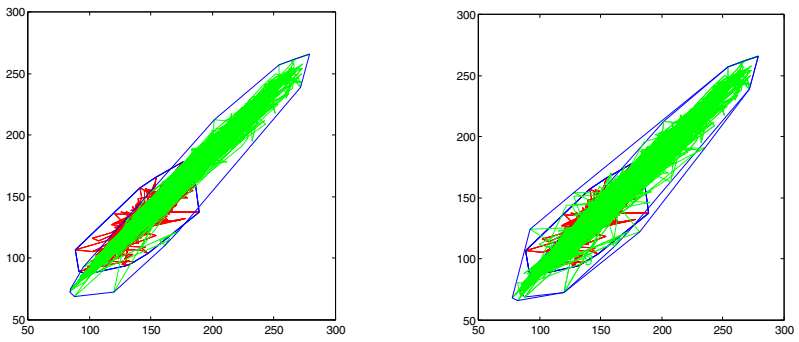

Fig. 7 The comparison of the data block of 0 am and 6 am~8 am. **a** The data block of 0 am and 8 am. **b** The data block of 0 am and the accumulating of 6 am~8 am

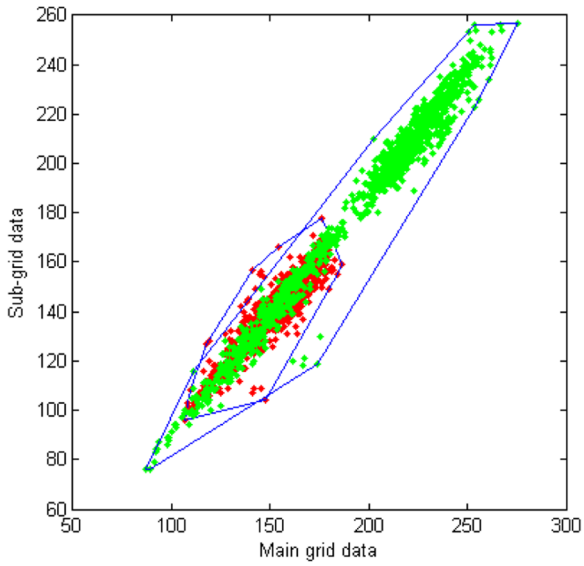

Fig. 8 The comparison of the new and old models

belongs to a new type of concept drift. Therefore, we use it to train new model and save it into the pool.

We also studied that when the drifting occurred, the correlation coefficient value of two blocks to be compared is less than or equal to 0.8 whose corresponding entropy value has a relatively large value. Data blocks could achieve accurate accumulate relatively, and when the sum of entropy is greater than a certain threshold, the system re-trains the model. Therefore, the feasibility of the updating mechanism proposed in this paper has been verified. Based on shape graphs of data blocks, when the edge shape and center changed obviously with the previous mode, the system would update the model and save it to the classifier pool. Through the management of the classifier pool, the system achieves the detection of concept drift more efficiently.

5 Conclusions

This paper proposed a mathematical model which is based on correlation coefficient and cumulative entropy calculation, introduced the significance of correlation coefficient and entropy and analysis the uncertainty of correlation, accumulated the sum of entropy value, and then make a judgment on whether to update the system or not. The model also designed a classifier pool, which saves the previous concept; the system could reduce the updating frequency and achieve updating efficiently. Finally, experimental results confirmed the validity of the proposed algorithm and verify the clustering system based on correlation coefficient and information entropy is an efficient mechanism.

Acknowledgements
This paper is supported by Natural Science Foundation of China (61271411) and Natural Youth Science Foundation of China (61501326, 61401310). It also supported by Tianjin Research Program of Application Foundation and Advanced Technology (15JCZDJC31500) and Tianjin Science Foundation (16JCYBJC16500).

Competing interests
The authors declare that they have no competing interests.

Received: 12 May 2016 Accepted: 26 August 2016
Published online: 13 September 2016

References

1. G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001: 97-106
2. ZW Yin, ST Huang, Adaptive method for handling local concept drift of data streams classification. Computer Science. **35**(2), 138–143 (2008)
3. Y Sun, GJ Mao, Mining concept drifts from data streams based on multi-classifiers. Acta Automat. Sin. **34**(1), 93–97 (2008)
4. HL Aboalsamh, A novel incremental approach for stream data mining. AEJ-Alexandria Engineering Journal **48**(4), 419–426 (2009)
5. LI Kuncheva, I Zliobaite, On the window size for classification in changing environments. Intelligent Data Analysis **13**(6), 861–872 (2009)
6. H Morshedlou, AA Barforoush, A new history based method to handle the recurring concept shifts in data streams. World Acad. Sci. Eng. Technol. **58**, 917–922 (2009)
7. I Katakis, G Tsoumakas, I Vlahavas, Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowl. Inf. Syst. **22**(3), 371–391 (2010)
8. YS Zhu, CM Guo, The research of correlation matching algorithm based on correlation coefficient. Signal Process. **6**, 007 (2003)
9. LL Minku, X Yao, DDD: a new ensemble approach for dealing with concept drift. IEEE Trans. Knowledge & Data Engineering. **24**(4), 619–633 (2012)
10. PD Zhou, *Computational geometry: algorithm design and analysis* (Tsinghua University press, Beijing, 2008)
11. H Borchani, P Larranaga, C Bielza, Classifying evolving data streams with partially labeled data. Intelligent Data Analysis **15**(5), 655–670 (2011)